# Exercise 3

(1)
In the JSON directory there is a CustomerSales.json file that contains customer details and all of the items they have purchased in their orders with the company. Have a look at the populate_customers\loadCustomerOrders.py file. When run, this script will load up the CustomerSales.json database into a collection called CustomerOrders

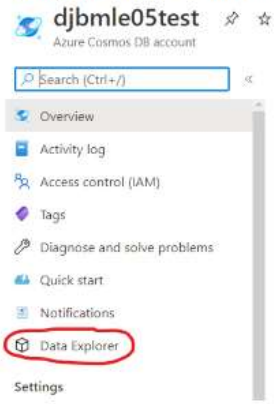
Examine the loadCustomerOrders.py file. Notice how the id field for each document is taken from the customer_id attribute in each document. Notice how we use the upsert_item method on the container object to load the data. Using this method makes the process idempotent as re-running the script will just replace all the existing document entries rather than duplicating them.

Run the loadCustomerOrders.py file using VSCode or from the command line, using the python conda environment we created in an earlier exercise

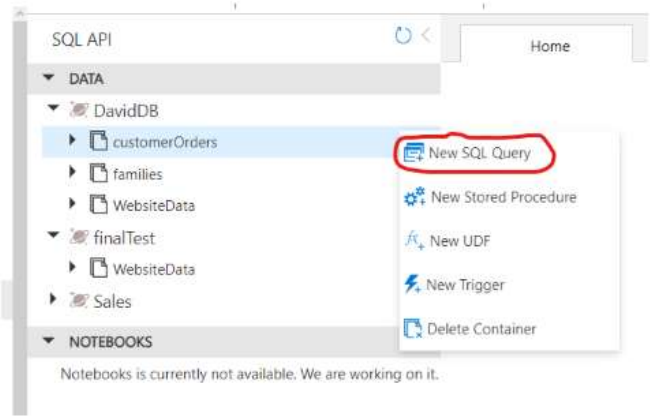
Python .\populate_containers\loadCustomerOrders.py

The script should take a few minutes to run .

(2)
Open up the Azure portal (portal.azure.com) and find the resource for your Cosmos DB Account.
Open up the Cosmos DB account overview page, locate Data Explorer in the left menu




Open Data Explorer, find the CustmerOrders collection and click on the ellipsis (...) on the right of the entry.




A new pane will appear with the SQL statement below

```
SELECT * FROM c
```

Modify the SQL statement to flatten out the customerOrders documents to return one document for every Order Item, but still containing the customer is, email, city and state. This statement will require you to use an intra document join using the JOIN statement.
See https://docs.microsoft.com/en-us/azure/cosmos-db/sql/sql-query-join if you are stuck.

Extend the statement you used to flatten out the data into Sales Order Items to determine the total sales by US State.
See https://docs.microsoft.com/en-us/azure/cosmos-db/sql/sql-query-group-by and
https://docs.microsoft.com/en-us/azure/cosmos-db/sql/sql-query-aggregate-sum if you need more help

Make a note of which state recorded the lowest sales.

(3)
In Visual Studio Code (or your favorite Python Editor) write a program to get the output of the query you created in (2) to get all of the customer orders with their customer details and store the output in a pandas dataframe.

Spend a few minutes working out how you will write this code. If you are really stuck ask the instructor for a hint.

Use the pandas dataframe you have created to calculate the sum of the values of all orders by US State. Check that is agrees with the results provided by you query in (2).

If there were 750 million records in the collection rather than a few thousand records how would your approach differ ?